Come on feel the noise - from metaphors to null models

Come on feel the noise - from metaphors to null models

Konrad Lohse[1]

[1] Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3FL, UK

The metaphor of "speciation islands" (Turner *et al.*, 2005) has dominated speciation research in the last decade. It invokes a particular (and plausible) feedback between divergent selection, migration and recombination (Barton & Bengtsson, 1986) and has led to a general re-evaluation of the role of gene-flow in speciation. This, together with the fall in sequencing costs, has spawned an industry of studies that scan the genomes of closely related species for outliers of divergence (usually measured in relative terms as F_{ST}). Undoubtedly, outlier scans have contributed to the discovery of spectacular examples of reproductive barrier loci in particular in *Heliconius* butterflies and cichlid fish (e.g Nadeau *et al.*, 2012; Malinsky *et al.*, 2015). Yet, given the flood of genomic studies, one may wonder why we have not learned more about the speciation process and the genes and genetic architectures involved in the build up of reproductive isolation.

Despite its appeal as a metaphor, the idea of speciation islands has proven frustratingly difficult to relate to sequence data in a concrete way. Studies invoking "speciation islands" as an explanation for outliers of divergence abound (Wolf & Ellegren, 2017), and a great deal of effort has been devoted to "follow up" on such outliers with genetic mapping or experimental studies. However, there have been few attempts to relate patterns of sequence diversity and divergence to the underlying population level processes in a quantitative way.

Ravinet *et al.* (2017) give a careful review of the demographic and selective processes involved in the build up of reproductive isolation and the complex ways they interact with each other to shape diversity and divergence along the genome. One of their main conclusions is that a meaningful interpretation of the genomic landscape of speciation must account for both the background demography as well as the heterogeneity in basic genome properties, such as gene density and the rate of recombination and mutation. Ravinet *et al.* (2017) also stress that divergence and diversity are highly stochastic and that incomplete lineage sorting (ILS) "[...] increases the variance of genomic divergence estimates making it difficult to identify true outliers and also potentially introducing false positives." Their conclusion, however, is oddly ambiguous: "[...] incorporating demographic history in tests for selection is difficult as incorrect specification of the history, potentially generated by ILS patterns increases error rates." They go on to argue that "[...] approaches that do not use demographic models may be preferable in some cases, although these too are prone to bias."

I agree about the difficulty of the task, but – contrary to Ravinet *et al.* (2017) – I would argue that model-based inference is the only hope for understanding the genome signatures of speciation. It is of course true that we are a long way from being able to fit a full, mechanistic model of the speciation process to genomic data. As Ravinet *et al.* (2017) make clear, even an extremely simplified cartoon of speciation at the genomic level must necessarily be complex and include demography, heterogeneity in background selection and recombination, specifics about the number and distribution of barrier loci and about how and when selection has acted on them.

Perhaps even more worryingly, we currently have no general understanding of how much information about past demography and selection is actually contained in genome data. In other words, even if we had a perfect method for extracting all the relevant signal, it is not clear how much detail we would be able to infer about a particular speciation history from genomic data alone. Clearly, the information in sequence variation is finite, while the space of potential speciation scenarios is not. The fact that even very simple

demographic histories for a single population have recently been shown to be non-identifiable from the site frequency spectrum (Terhorst & Song, 2015; Lapierre *et al.*, 2017) is a pertinent reminder of the inherent limits of the information in sequence data. Thus, Ravinet *et al.* (2017) are right to emphasise the value of incorporating independent information in the form of recombination and background selection maps.

The idea that all the demographic and selective processes that shape a particular genomic landscape of speciation could be captured in a single model is daunting at best and infeasible at worst. However, a much more realistic and very worthwhile starting point for speciation genomics would be to ask how well a particular genomic landscape of can be explained by simple null models. Thanks to algorithmic improvements in coalescent simulations (Kelleher *et al.*, 2016), we now have – for the first time – the ability to efficiently generate genomic landscapes of divergence under the full ancestral process of coalescence and recombination (and even condition such simulations on a recombination map) and for any demographic scenario. To give a concrete example, consider the simplest possible null model of speciation: a strictly allopatric split without any selection or heterogeneity in recombination. Coalescent simulations under this model give an immediate feel for the noise inherent in divergence and diversity measures (Fig. 1). Importantly, it is clear that more or less pronounced but entirely random F_{ST} peaks arise as a result of genetic drift even under under this simple history. This is especially true if we allow for population structure within species (Fig. 1 bottom).

While the effects of background and positive selection on measures of divergence have been amply pointed out (Charlesworth, 1998; Noor & Bennett, 2009; Cruickshank & Hahn, 2014), the challenge that the randomness of the coalescent poses for outlier scans is often ignored. Part of the reason may be that theory and simulation studies on selection during speciation tend to focus on average divergence and diversity (e.g. Feder & Nosil, 2010; Guerrero *et al.*, 2011; Aeschbacher *et al.*, 2016). For example, Yeaman *et al.* (2016) study the effect of selection on a locally beneficial variant that is linked to an existing polymorphism under

migration-selection balance in terms of *average* F_{ST}. Although *mean* divergence is the natural starting point for theoretic work, in order to make sense of genomic data, we need to know the *distribution* of divergence and diversity both in the presence and absence of barrier loci (Lohse *et al.*, 2016).

The fact that the majority of scans for divergence outliers have been agnostic about the underlying demography and have defined outliers simply as the extreme tails of the divergence distribution (Wolf & Ellegren, 2017), means that we currently have no idea what fraction of the "significant" F_{ST} peaks reported simply reflect the randomness of the coalescent. A straightforward way to distinguish variation in neutral divergence from signal of selection on barrier loci is to obtain significance thresholds of divergence estimates such as F_{ST} from coalescent simulations that condition on a specific background history. This approach acknowledges that the power to detect divergent selection on barrier loci is a function of both the demographic history and the relative rates of and heterogeneity in recombination and mutation along the genome. For example, we expect less heterogeneity in F_{ST} under a history of divergence with gene flow as simulated by Ravinet *et al.* (2017), while population structure greatly increases the heterogeneity in F_{ST} (Figure 1, bottom). Similarly, while the recombination and mutation rates assumed in the simulation in figure 1 were based on estimates in *Drosophila* (Keightley *et al.*, 2009), we would expect much noisier F_{ST} trajectories for taxa with a relatively lower recombination rate (i.e. ρ/μ). Conditioning outlier scans on explicit demographies has several immediate benefits: First, it becomes possible to diagnose genomic landscapes that do *not* contain any detectable barrier loci either because there are none (i.e. divergence occurred in allopatry) or because there are too many (i.e. RI traits are highly polygenic) or both, yet may still look "peaky" (Fig. 1). Second, we can distinguish speciation histories that involved ongoing gene flow from isolation followed by secondary gene flow. As Ravinet *et al.* (2017) point out, these can give rise to very similar genomic landscapes. Third, we can focus efforts on those speciation histories that are most interesting (i.e. involving gene flow) and informative about the selective events during speciation. Finally, demographically explicit

genomic landscapes will allow for much more meaningful comparisons across taxa which are essential if we want to draw general conclusions about how speciation happens.

The plea for model based inference is of course not an argument against the usefulness of simple and intuitive summary statistics for visualising and exploring genomic data. However, if we want to interpret the genomic landscapes of speciation in terms of the underlying processes, we have no choice but to model those processes. Models force us to be explicit and, given the efficient simulation tools now available (Kelleher *et al.*, 2016; Haller & Messer, 2017), can be easily confronted with data. Unlike metaphors, the purpose of models is that they can (and should) be updated if they turn out to be no good.

Acknowledgements

I would like to thank Crispin Jordan for useful discussions and comments on this note and Mike Ritchie for the invitation to write it. KL is supported by an Independent Research fellowship from the Natural Environment Research Council (NE/L011522/1).

References

Aeschbacher, S., Selby, J.P., Willis, J.H. & Coop, G. (2016). Population-genomic inference of the strength and timing of selection against gene flow. *bioRxiv*. doi:10.1101/072736.

Barton, N. & Bengtsson, B.O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57(3), 357–376.

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, 15(5), 538–543.

Cruickshank, T.E. & Hahn, M.W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23, 3133–3157. ISSN 1365-294X. doi:10.1111/mec.12796.

Feder, J.L. & Nosil, P. (2010). The efficacy of divergence hitchhiking in genering genomic islands during ecological speciation. *Evolution*, 64(6), 1729–1747. ISSN 1558-5646. doi:10.1111/j.1558-5646.2009.00943.x.

Guerrero, R.F., Rousset, F. & Kirkpatrick, M. (2011). Coalescent patterns for chromosomal inversions in divergent populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1587), 430–438. ISSN 0962-8436. doi:10.1098/rstb.2011.0246.

Haller, B.C. & Messer, P.W. (2017). Slim 2: Flexible, interactive forward genetic simulations. *Molecular Biology and Evolution*, 34(1), 230. doi:10.1093/molbev/msw211.

Keightley, P.D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S. & Blaxter, M.L. (2009). Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research*, 19(7), 1195–1201.

Kelleher, J., Etheridge, A. & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, 12(5), 1–22. doi:10.1371/journal.pcbi.1004842.

Lapierre, M., Lambert, A. & Achaz, G. (2017). Accuracy of demographic inferences from site frequency spectrum: The case of the Yoruba population. *bioRxiv*. doi:10.1101/078618.

Lohse, K., Chmelik, M., Martin, S.H. & Barton, N.H. (2016). Strategies for calculating blockwise likelihoods under the coalescent. *Genetics*, 2(202), 775–786.

Malinsky, M., Challis, R.J., Tyers, A.M., Schiffels, S., Terai, Y., Ngatunga, B.P., Miska, E.A., Durbin, R., Genner, M.J. & Turner, G.F. (2015). Genomic islands of speciation separate cichlid ecomorphs in an east african crater lake. *Science*, 350(6267), 1493–1498. ISSN 0036-8075. doi:10.1126/science.aac9927.

Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., Quail, M.A., Joron, M., ffrench Constant, R.H., Blaxter, M.L., Mallet, J. & Jiggins, C.D. (2012). Genomic islands of divergence in hybridizing heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1587), 343–353. ISSN 0962-8436. doi: 10.1098/rstb.2011.0246.

Noor, M. & Bennett, S. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, 103(6), 439–44.

Ravinet, M., Faria, R., Butlin, R., Galindo, J., Bierne, N., Rafajlovic, M., Noor, M., Mehlig, B. & Westram, A. (2017). Interpreting the genomic landscape of speciation: road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, 5(10), e1000695.

Terhorst, J. & Song, Y.S. (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences*, 112(25), 7677–7682.

Turner, T.L., Hahn, M.W. & Nuzhdin, S.V. (2005). Genomic islands of speciation in *Anopheles gambiae*. *PLOS Biology*, 3(9). doi:10.1371/journal.pbio.0030285.

Wolf, J. & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nat Rev Genet*, 18(2), 87–100.

Yeaman, S., Aeschbacher, S. & Bürger, R. (2016). The evolution of genomic islands by increased establishment probability of linked alleles. *Molecular Ecology*, 25(11), 2542–2558. ISSN 1365-294X. doi:10.1111/mec.13611.

Figures

Figure 1 F_{ST} along a 8.5 Mb stretch of sequence simulated under a history of strict divergence without gene flow at time $T = 0.5$ (measured in $2N_e$ generations). Parameters were motivated by *Drosophila*: $N_e = 0.5 \times 10^6$, $\rho = 1.15 \times 10^{-8}$ and $\mu = 3.46 \times 10^{-9}$ per base and generation (Keightley *et al.*, 2009). F_{ST} computed in 15kb and 37.5kb sliding windows (dashed and solid lines respectively) varies substantially around its mean (gray line). Pronounced, but entirely random peaks in F_{ST} arise in particular for small samples ($n = 4$) (top) and when populations are structured within each species (bottom) (a sample of $n = 10$ taken from 1 of 10 demes connected by symmetric migration at rate $M = 4N_d m = 0.8$).

Figure 1: F_{ST} under a null model of allopatric divergence.